Demystifying Data Science Projects: A Look on the People and Process of Data Science Today

Aho, Timo

Springer
2020

Aho , T , Sievi-Korte , O , Kilamo , T , Yaman , S & Mikkonen , T 2020 , Demystifying Data Science Projects: A Look on the People and Process of Data Science Today . in M Morisio , M Torchiano & A Jedlitschka (eds) , International Conference on Product-Focused Software Process Improvement . Lecture Notes in Computer Science , vol. 12562 , Springer , Cham , pp. 153-167 , International Conference on Product-Focused Software Process Improvement , Turin , Italy , 26/11/2020 . https://doi.org/10.1007/978-3-030-64148-1_10

http://hdl.handle.net/10138/322911
https://doi.org/10.1007/978-3-030-64148-1_10

other

Demystifying Data Science Projects: a Look on the People and Process of Data Science Today

Timo Aho[1], Outi Sievi-Korte[2], Terhi Kilamo[2], Sezin Yaman[3], and Tommi Mikkonen[4]

[1] TietoEVRY, Tampere, Finland
timo.aho@iki.fi
[2] Tampere University, Tampere, Finland
outi.sievi-korte@tuni.fi, terhi.kilamo@tuni.fi
[3] KPMG Finland, Helsinki, Finland
sezin.yaman@kpmg.fi
[4] University of Helsinki, Helsinki, Finland
tommi.mikkonen@helsinki.fi

Abstract. Processes and practices used in data science projects have been reshaping especially over the last decade. These are different from their software engineering counterparts. However, to a large extent, data science relies on software, and, once taken to use, the results of a data science project are often embedded in software context. Hence, seeking synergy between software engineering and data science might open promising avenues. However, while there are various studies on data science workflows and data science project teams, there have been no attempts to combine these two very interlinked aspects. Furthermore, existing studies usually focus on practices within one company. Our study will fill these gaps with a multi-company case study, concentrating both on the roles found in data science project teams as well as the process. In this paper, we have studied a number of practicing data scientists to understand a typical process flow for a data science project. In addition, we studied the involved roles and the teamwork that would take place in the data context. Our analysis revealed three main elements of data science projects: Experimentation, Development Approach, and Multidisciplinary team(work). These key concepts are further broken down to 13 different sub-themes in total. The found themes pinpoint critical elements and challenges found in data science projects, which are still often done in an ad-hoc fashion. Finally, we compare the results with modern software development to analyse how good a match there is.

Keywords: Data science, data engineering, software process, prototyping, case study

1 Introduction

The layman's view to a data science project is glorified, to the brink of data scientists being modern-day fortune tellers, seemingly effortlessly creating predictions

based on existing data. The reality, however, is somewhat different. While the final outcomes of a data science project can appear miraculous, the actual data science – as well as related activities such as data engineering and data mining – build on well-established ground rules on what the data says and what it does not say.

The terminology in the field of data science is somewhat mixed, with overlapping terms like data analytics, machine learning, data mining and big data. In this study, we use the term data science for extracting knowledge from data sets, which might be large, using multidisciplinary techniques such as statistics and machine learning, in order to understand and analyze the data and to gain insights. However, here we exclude traditional business intelligence and data warehousing from the scope of data science.

Today's data science projects exhibit some problems that could be tackled with more mature project management methodologies [8, 9]. These include minimal focus on identifying result quality and problems in estimating budget and scheduling in advance [18]. In addition, since many of the data science results are applied in the context of software systems, seeking synergy between software development approaches and data science seems to open promising avenues. For instance, Sculley et al. [23] state that for a mature machine learning system, it could be that only at most 5% of the overall code base can be regarded as machine learning, a subset of data science. Rest of the code is about, e.g., data collection and preparation, configuration and management, and serving layer. This raises the question, whether following readily available approaches in software development could help in data science projects [1, 22].

In this paper, our goal is to understand a typical process flow for a data science project, as well as to learn about the role of a data scientist and teamwork that would take place in the context of data-centric projects. Our precise research questions are:

1. What is the typical process flow of a data science project?
2. What kind of people are part of a data science project?

The research was executed as a multiple case study with a series of interviews with experienced data scientists working in the field of data science consultancy.

Our results indicate that data science is experimentation-centric and multidisciplinary team work. The role of a data scientist is identified as distinctively separate from that of a data engineer. Development is mainly iterative in nature. As the work relies heavily on experimentation on data, models, algorithms and technical approaches utilized, knowledge gained during the project can change goals or requirements of the work. However, in the context of larger projects, practices sharing characteristics with modern software development are common, in particular when the team size increases.

The rest of this paper is structured as follows. Section 2 gives the background for the paper and presents related work. Section 3 introduces the research approach we have followed, and Section 4 presents the results of the study. Section 5 provides an extended discussion regarding the results, including also threats to validity. Finally, Section 6 concludes the paper with some final remarks.

2 Background and Related work

Typically, data science related research concentrates on the technological solutions and their use cases as presented in the survey by Safhi et al. [16]. At the same time, literature on data science project roles and project methodologies is scarce, while there has been some growth in the field [21].

So far, data science projects have been following their own processes and practices, which have been different from those that have been typically used in the context of software development [13, 19]. Data science specific project methodologies include KDD [4], CRISP-DM [26] and SEMMA[5]. Of these, CRISP-DM seems to be the one most referred to. It describes an iterative process with six stages: *a*) business understanding, *b*) data understanding, *c*) data preparation, *d*) modeling, *e*) evaluation, and *f*) deployment. The stages follow one other linearly, but the process allows both moving back and forth between the stages. For a comparison of the frameworks, we refer to the work of Shafique and Qaier [24], and Azevedo and Santos [3]. There are also extensions (e.g. [2, 7]) on these methodologies that aim at tackling some of the problems the practitioners have identified.

In an older 2015 Internet poll [13], CRISP-DM was shown to be the most popular process methodology in data science. However, according to a more recent 2018 survey by Salz et al. [19], 82% of data science teams did not follow any explicit project management process or practices, even though 85% thought such would be beneficial. According to the survey, teams either were not sure of the used process methodology, or used an ad hoc approach. Moreover, 15% of teams reported the use of some agile methodology and 3% a CRISP-DM based methodology.

Grady et al. [8] note the similarity of data science projects with software development before the adoption of agile methodologies. Such similarities can also be seen in a study revealing difficulties related to processes in data science projects [18]. Issues were found particularly with estimating the budget, schedules and the successfulness of the project in advance. Also quality assurance of results is often insufficient. Moreover, data science projects still rely quite heavily on individual effort instead of team work.

However, it is important to note that there are differing categories of data science projects. For example, in an ethnographic study [20] the authors found two kinds of data science projects: routine data transformation projects and exploratory projects. Especially the latter ones were one-off and did not have standard process methodologies in use; for example, the projects lacked milestones and schedules. Moreover, the time used for different stages varied a lot and included a lot of manual work on, e.g., data transformations.

In a further study, Saltz et al. [17] could label data science projects with two dimensions (infrastructure and discovery), based on which they could identify four different types of data science projects depending on where projects could

[5] Available at `https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=15.1`

be placed on the axes. The project types were: Hard to Justify, Exploratory, Well-Defined and Small data.

Similarly, Amershi et al. [1] discuss how data science teams at Microsoft have merged their data science workflows into pre-existing agile processes. While data scientists have learned to adapt to the agile workflow, the authors recognize several issues related to data, model creation and handling AI components, that clearly distinguish data science projects from software projects. The authors note, though, that the problems faced by data scientist change significantly based on the maturity of the team, and have created a maturity measure to help identify issues.

The nature of data science teams and member backgrounds have also been studied. Kim et al. [11, 12] identify that data scientists could have very different kinds of roles in teams and projects, partially due to their interest, skills and background, and partially due to company principles on how work is divided. Data scientist profiles vary from "Polymath" who has a strong mathematical background and can handle technical implementation, to "Insight actor" whose main job is to act based on findings from the data.

In general, most of the studies concentrate on a single company or are structured surveys with large target groups. There are only few (e.g. [10, 17]) data science interview studies over multiple companies. Kandel et al. [10] concentrate on individual analyst skill set and workflow mentioning within team collaboration briefly. Moreover, Saltz et al. [17] give a data science project framework mentioning management and organization as a social context.

To summarize, prior work on data science projects investigates software development approaches and highlights the parallels and differences. However, to the best of our knowledge, no current research across multiple data science companies exists. Further, there are studies on different workflows and types of data science projects, and also studies on what kind of teams are used within data science projects. Nevertheless, to the best of our knowledge, no study yet exists that would combine these to angles together. Our paper attempts to fill this gap.

3 Research Methodology

The goal of this work is to understand a typical process flow for a data science project, and to learn about the role of teamwork that would take place in the context of data science projects, and what is the role of the data scientist there. The study was conducted as a multiple-case study of six companies with a business area in data science consultancy (interview protocol is online[6]). Case study research [27, 15] as an approach is suitable when the aim is to gain knowledge on a topic tied to and not clearly separable from its practical context. This is

[6] The interview protocol
`https://drive.google.com/file/d/1rKvt_1OoeINvOhXvQUQHgIgtFyEj9sAf/view?usp=sharing`

true for industrial data science projects where practitioners can provide a good view on how everyday data science work is done today.

The interview questions were iteratively designed by the authors, taking into consideration existing related work and some baseline assumptions. We identified five assumptions based on prior research and our own observations from industrial experience – authors 1 and 4 are currently working as data scientists where as author 5 has extensive experience in industrial software development. The assumptions driving the focus of the work were the following:

- Data scientists are lone warriors or miracle workers, who come in to do a data science element and then leave after a short time, never seeing the project complete and never being truly part of the development team.
- Broken data presents challenges to data science work.
- Insufficient data presents challenges: clients' needs can not be met because there is no available data to answer the clients' targets.
- Data science projects are vaguely specified and customers do not exactly know what they want in the beginning of the project.
- Data engineering and data science are clearly separated tasks.

Note that the assumptions are not hypotheses, but are included for the sake of openness and validity.

Six data science consultancy companies were selected into the study based on availability and the nature of data science projects they work with. Three of the companies were general ICT consultancy companies with roughly 500–1000 personnel. The other three focus specifically on AI, data analysis and concept design. Two in the latter group were independent companies with less than 50 employees. One was a data science unit of a similar size within a large, global business consultancy company.

Table 1. Data Science experience of the interviewees in years.

Experience type	Experience in years					
Data Science Consultancy	2	4	4	7	7	12
Overall Data Science	NA	9	NA	13	21	12

An experienced data scientist was interviewed from each company (see Table 1). The interviews concentrated on overall experience of the data scientists over their whole career. Thus, interview questions did not address, e.g., the related project details. The interviews were conducted from November to December 2019 as a semi-structured interview lasting approximately half an hour. The interview protocol was designed based on the assumptions, and the first interview acted as a pilot interview for the interview protocol. As no changes were needed after the pilot, the pilot interview is included in the analysis. Five of the interviews were done on the companies' premises and one on a university campus. All interviews were done in the native language of the interviewees. Two researchers

were present in each of the interviews, one of them taking notes. Each interview was recorded and transcribed.

The used definition of a data science project was given in the beginning of each interview but further specifics were left to each interviewee. In the scope of the study, a data science project must apply programming and not just use graphical tools in the analysis. Furthermore, the project has to include artificial intelligence or data science development, for instance predictive or exploratory analytics. It was also emphasized that traditional business intelligence or data warehousing were not within the scope of the study.

The results were thematically analyzed [5] based on the notes and the transcriptions. One researcher made the thematic analysis based on the notes and the transcripts, arriving at three higher level themes which comprised of 13 lower-level themes in total. Once an initial thematic analysis was made, another researcher validated the analysis by placing 20 quotes (chosen randomly but in such a way that all themes were represented) under themes identified. Once the themes were agreed upon by two researchers, their analysis was further validated by a third researcher in the same way. The coding essentially remained the same after validation, and no changes were made to the lower-level themes. However, two higher-level themes were named more appropriately, and some re-arranging was done in how lower-level themes were grouped under the higher-level themes. The themes are described in the following section.

4 Interviews

Based on our thematic analysis, we created a conceptual model of key elements encountered in a data science project (Figure 1). The three main concepts are *Experimentation*, *Development Approach*, and *Multidisciplinary Team(work)*, which we will present in more detail in the following.

4.1 Experimentation

Data science projects revolve around experimentation and dealing with the uncertainty of unpredictable outcomes. Data scientists need to experiment with data, models, algorithms, and technical approaches to find the most satisfying way of meeting their goals. Knowledge gained during the experimentation phase may lead to changes in goals or requirements, to more accurate models, and eventually to a Proof-of-Concept implementation.

Data — Based on our own experiences, we approached the interviews with an assumption that incomplete or broken data would present significant challenges in data science projects. All our interviewees agreed that data is never perfect: it is often flawed and incomplete, and has far less information value than what customers usually believe. It is accepted as status quo that you simply need to invest the necessary time to fix and clean the data. However, contrary to our assumptions, this was not considered to be a particular challenge, as it is something that data scientists come against in virtually every project.

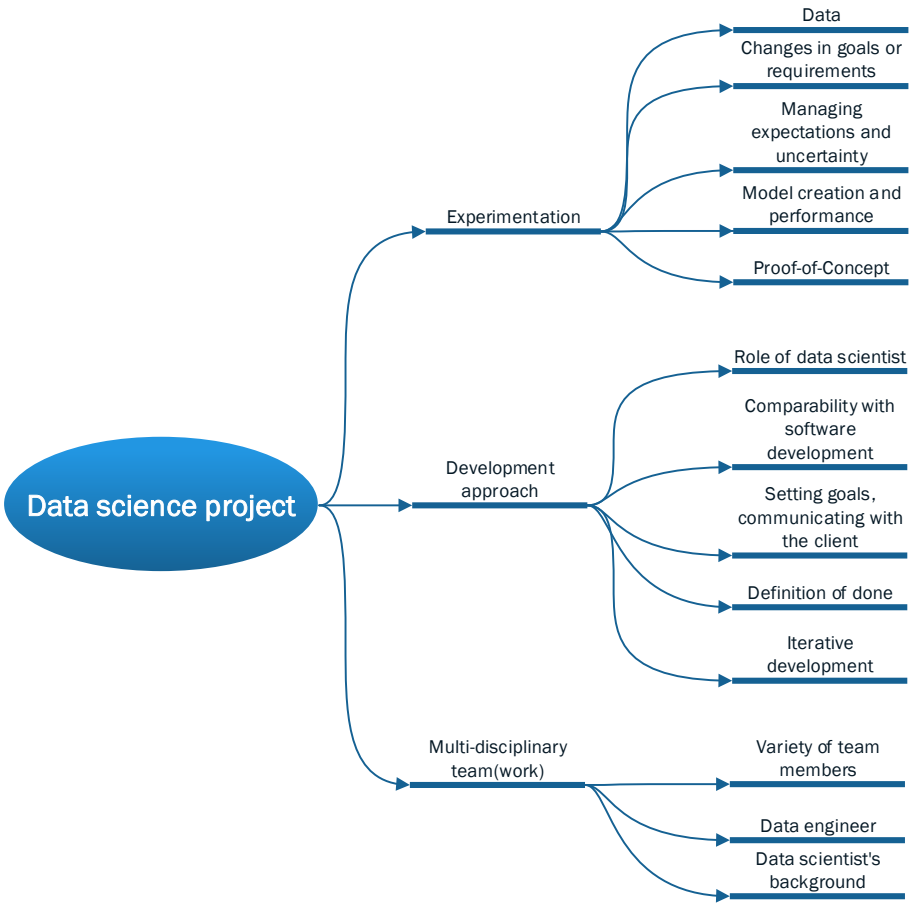

Fig. 1. Conceptual model of Data Science projects

Changes in Goals or Requirements — Due to the experimental nature of data science projects, goals or requirements often change over time. One reason is the aforementioned incompleteness of data: once the data scientist has done a first look through and created an understanding of the data, it may become apparent that the original goals simply are not feasible with the existing data. Our interviewees had a consensus that on a larger scale the goals and requirements for a data science project stay the same, but on a smaller scale the details may change based on the knowledge gathered through experimentation. Another setup is that new goals are created to complement the original ones. *"It's data* ***science*** *– you don't know in advance what can be achieved, chances may be improved with experience. It's common that goals slightly change."*

Managing Expectations and Uncertainty — The strongest message conveyed by our interviewees was that in data science one never really knows the outcome. This unfolds as significant challenges when communicating with the client and specifying the project. Our interviewees report that the present hype around AI is making things even more difficult. Various companies are reacting to the hype and on stories how data science projects have created value for others. However, the level of maturity for data collection and understanding the boundaries of data science varies significantly. Data scientists feel pressed in keeping expectations realistic and clearly communicating that one simply cannot know what, e.g., a machine learning algorithm will actually learn from the given data: *"Results aren't certain. If you promise too much, you are facing a difficult project. You have to be honest."*

Model Creation and Performance — The core of data science is creating models and utilizing algorithms to generate information value from the data. Our interviewees discussed various ways of conducting data science: doing reinforcement learning, "simple" machine learning, data exploration, predictive modeling, and natural language processing. However, they also raised some issues: 1) exploratory approaches may be problematic, if data is separated from the context, 2) having a model that somehow works with the data is only 5% of the project done, and 3) models are a product of iterative improvement and their performance could be honed forever.

Proof-of-Concept — As described, data science projects have a high level of uncertainty, an experimental and iterative nature of developing models for data, and an increased need to have something tangible to discuss goals with the client. Hence, it appears natural that the projects are heavily reliant on creating Proof-of-Concept (PoC) implementations. Our interviewees unanimously discussed situations where the feasibility of creating a model for the given data and making some sensible results from it were tested by creating a PoC. During the PoC development there may only be one or two data scientists involved and no other team members. The whole PoC would be developed in only a few week's of time. *"We test a little and do PoC versions of what we are planning on developing. We get some certainty that our approach makes sense."*

4.2 Development Approach

The development approach in data science projects seems to incorporate data scientists into larger development teams. The work is also clearly iterative in its nature, as iterative development approach was applied according to five out of the six interviews. Furthermore, a parallel with software development was drawn in four of the six cases. However, there can be significant differences in goal setting based on the maturity of the client.

Role of Data Scientist — Our assumption prior to the interviews was that data scientists' role is solo work, where they only come in to complete the data science element never seeing the project complete and never truly working as a part of a development team. This turned out not to be true. While some data scientists worked on data with a clear cut focus on data science work,

there was strong commitment to the overall project. The work effort varied from sharing commitment between jobs to full commitment to one project. This depends on the stage of the project and the need for data scientist in the project. Some data scientists also work on data engineering, but the role of the data engineer is overall recognized separately (see Section 4.3). One interviewee also raised the topic of client contact. Data scientists need understanding of the client organization's needs to be able to provide data science solutions to meet those needs.

Comparability with Software Development — A lot of parallels were seen between data science work and general software development. There was a drive to get data science work to follow the process approaches commonplace in software development. Also the data science component was mentioned as just a small piece in a far larger project. One interviewee: *"It comes probably as a surprise to many what I mentioned earlier that you have 5% of machine learning and 95% of something else"*. The ending of the project was seen different from software development in that in data science the project was mentioned to never finish. Instead, only time or budget constraints determined the end of the data science project.

Setting Goals, Communicating with the Client — We assumed that data science projects are vaguely specified and customers do not have a clear goal in mind when the project starts. According to the interviews, the data science experience of the client was seen as a key factor in setting the project's goals. Business goals were mentioned as being at the heart of project goal setting. While having clear goals at the start of the project was considered valuable compared to starting with "what can you find out from the data", how well such goals are defined were seen to depend on the maturity level of the customer. However, the quality of goals set was considered to have begun to go down as data science has become more widely utilized. Communication with the client requires, in addition to having understanding of the clients business, the ability to set the expectations to a suitable level in order to meet the goals set.

Definition of Done — Definition of done in data science projects was twofold. Firstly, the maintenance phase can act as a clear ending to the project. Once the data science component is validated and in production, the work is done. Secondly, there are projects that go on indefinitely. There is always room for improvement, such as model calibration, and the project either goes on with new improvements as long as it is funded or spawns new projects to continue the work in.

Iterative Development — Based on the interviews, there is a clear link between the iterative nature of development and experimentation. Still, when viewed as a theme of its own, the role of the iterative development approach is clearly seen. Data and algorithm selection require iteration to find the most suitable solutions. Also having unclear goals requires iterations to make it possible to see what can be achieved. In some cases a specific approach for development was utilized: Scrum was mentioned as well as the use of sprints. When specified, the length of a sprint was one to two weeks.

4.3 Multi-Disciplinary Team(work)

While data science projects can be executed as small PoC efforts, where only one or two data scientists use a week's effort in testing an idea, teamwork is required when the data science modules are used in production. If a PoC is successful in demonstrating the feasibility of an idea, it can be utilized in a larger concept – a software product, automation, or another domain. In this setting a larger team with varied expertise and roles is required with different roles in, e.g., sales, marketing, and software development, to complement the data science and engineering skills.

Variety of Team Members — As the interviews revealed, data science can be a small part of a larger development project. When the data science component needs to be integrated with other components, a variety of team members are required for the project to succeed. Our interviewees stressed the role of software developers, to quote: *"A software developer is really important, and a good developer will save you from the trouble you didn't know you'd get into."* Developers' expertise varies similarly as in any regular software development project. Additionally, the interviewees stressed the importance of having someone who can understand the business side of things and fluently communicate with the customer. As noted, managing expectations, dealing with uncertainty and setting goals requires some special effort in data science projects, and, thus, team members with communication skills and business understanding are highly valued.

Data Engineer — As discussed, data is far from perfect in terms of being usable for data scientists. Also our initial assumption was that engineering the data is a clearly separated task, preceding the actual data *science*. Before a data scientist can begin, a working technical pipeline is required to actually access and gather data. Data may need to be fetched from several big databases, it can be in different formats or encrypted, and it can be as big as millions of rows, which requires partitioning. Performing such tasks requires an understanding of the client's data warehouse, as well as strong skills in database design and programming. After required data is gathered from the clients and put in a system where data science tasks will be performed, the data still requires polishing and fixing before it can be used with a model. Our interviews revealed that engineering the data is definitely considered to be distinctively separate from data *science*, and the role of data engineer was unanimously recognized in the interviews. However, who adopts the role of data engineer varies a great deal. In some cases the data scientists do data engineering as well, in some cases a software developer takes on the role of data engineer, and finally there may be people distinctively assigned the role of data engineer. Assignment of the role depends on multiple aspects: the scope and nature of the project and the data, the policies and practices of the company, and the backgrounds and profiles of data scientists. While having developers do the work of data engineer is quite in line with our initial assumptions, cases where a data scientist does data engineering as well is clearly contradictory, and we would need to probe further into defining the scope of data engineering that a data scientist actually does.

Data Scientist's Background — While our interviewees agree that data scientists can come from various backgrounds, they further agree on a common denominator: the ability to understand mathematics, understand, design and implement algorithms, and quickly learn new methods. Data scientists have a variety of educational backgrounds, such as economics, mathematics and physics. In addition, there are a large number with a technical background and even a tailored doctoral degree in machine learning. Finally, there may be so-called "Full stack data scientists", who are able to engineer the data, create the model, implement the algorithms, and also develop the software surrounding the data science module. While various backgrounds give sufficient skills in working as a data scientist, naturally the background affects how the data scientist approaches a problem. To quote one of our interviewees: *"a mathematician looks at the world completely differently from a statistician"*.

5 Discussion

Next, we present our analysis of key findings and list some key lessons learned. Then, we address validity concerns of our study.

5.1 Process and Collaboration in Data Science

The main concepts that define the process and roles of data science projects are *Experimentation*, relying on an iterative *Development Approach* and especially larger projects having *Multidisciplinary Team(work)* at their core.

The exploratory nature of many data science projects is evidenced by the proposed models and methods which have built-in learning mechanisms. Furthermore, some characteristics of prototyping [6] can be considered compatible in the context of data science. Proof-of-Concepts done by one or two data scientists appear to be a common mechanism to test the feasibility of a solution, do some initial testing of a model, and get familiar with the data. However, they are rarely sufficient as such.

Our results also reveal that there is a need for larger data science projects where numerous team members participate in different roles. These larger projects can also involve a considerable amount of software development, where data related features are embedded. This is also reflected in team composition. The interviewees state that *"In many cases there are 1–2 data scientists, then a varying number of people in software development"*. Here data scientists are also involved throughout the project, all the way to the maintenance phase.

Based on our results, the *development approach* of data science projects appears to rely at the heart of iterative work – a process familiar in the context of software development. Five out of six of our interviewees commented on the iterative role of the development, and the final interviewee referred to parallels of software development in general. Nevertheless, even in software development, there are multitude of ways of iterating for different purposes [25]. Thus, this

shows that there are similarities between data science projects and software development, but this does not necessarily mean that the two could be aligned easily. Data science projects come with an exceptionally high level of uncertainty on the outcome, as was revealed both by our study and in related work [19]. Within data science, that uncertainty is acknowledged and accepted to a point – data scientists are well aware that one cannot know what can be derived from the data before experimentation. However, that uncertainty stretches from the start of the project (vague specifications) to the end (very varied conceptions of what is the "definition of done"), and may be very difficult to accept when moving to a software development context.

Drawing our findings together indicates that data science projects can benefit from development processes of software development, especially in larger projects. Based on the experimentative nature of data science work one can argue that what is commonly called a Proof-of-Concept implementation in data science could probably be regarded as a prototype in software terminology. Team work is at heart of data science work – both between several data scientists as well as between the data scientist and other, often software, professionals. These elements of working up iteratively from Proof-of-Concepts and prototypes and forming multidisciplinary teams for development are commonplace in agile and lean software development practices.

5.2 Threats to Validity

Several threats to validity [15, 27] are recognized. Mitigation factors are also taken into account. For the study, we especially address construct validity, external validity and the reliability of the work.

Construct validity — Construct validity considers how well research investigates what it means to investigate. In this study, construct validity is threatened by how representative the interviewed cases were of data science and how the interview data was analyzed. Also the case selection was partly based on a convenience sample. To mitigate the threat, what was meant by a data science project was defined in the beginning of each interview. However, the definition given left room for further specifics by each interviewee. Furthermore, the interviewees were selected based on their prior experience in data science projects specific to the scope of the study. All researchers participated in the planning and development of the interview protocol, which was also piloted in a pilot interview.

External validity — External validity refers to how well the study results can be generalized beyond the scope of the study. This study has been planned as a preliminary to a larger survey study, that is currently being designed. While the sample size of this study is limited, the interviewees were selected to represent a range of experience and from two fields of data science industry (general ICT and artificial intelligence specific external consultancy companies) with small to medium company sizes. Our findings may not be applicable to in-house teams, nor to larger enterprises or smaller companies considering the sample size. However, we believe that the study results give important insights toward under-

standing data science projects and developing a theory. As future work, we aim at validating these results.

Reliability — The main threats to the reliability of the results is in the thematic analysis. To mitigate the threat, three researchers took part in the interviews and all five participated in the analysis of the results. The main thematic analysis was done by one researcher and was validated by two researchers separately. The conflicts that occurred during the validation were resolved in a separate analysis sessions with the rest of the researchers' participation. This way, researcher bias was also minimized, as three of the authors had prior experience with regard to data science and software engineering in practise. To enable to replication of the study, the interview protocol is available online.

6 Conclusions

In this multiple case study, we interviewed six data scientists with different levels of experience from six small to medium-sized consultancy companies. Our aim was to understand a typical process flow for a data science project, as well as to learn about the role of teamwork and data scientists there.

Three main concepts describing data science project methodology and roles were found. *Experimentation* is a core nature of data science projects. The data science projects commonly have an iterative *Development Approach* that incorporates them into larger teams. In addition, successful Proof-of-Concepts (PoCs) often end up in larger projects having *Multidisciplinary Team(work).*

Our first research question was: *1. What is the typical process flow of a data science project?* From the project perspective we found that process elements in data science projects were, to some extent, the same as in software development. However, what sets the data science projects clearly apart from software development is the inherent uncertainty of data science work. It must be acknowledged and clearly communicated that there is no guarantee of specific results or achieving the initial goals. At the same time, requirements in software development are also fuzzy even at best. Nevertheless, software development has processes and tools to handle the uncertainty whereas data science, in turn, has to live with the data being inherently broken. Furthermore, the uncertainty cuts through the entire data science project life cycle – from vague specifications to differences in the definition of done. Data science projects are at core about experimentation and exploration, making them somewhat similar to the Lean Startup [14] cycle of Build – Measure – Learn.

To answer our second research question, *2. What kind of people are part of a data science project?*, the assumption of data scientists as lone soldiers was debunked. Firstly, data science is often only a small part of a larger development project. As the projects can be long ranging, there is commitment to the project throughout its life cycle. Especially, when a PoC demonstrates a feasible idea, a larger team with varied expertise and roles is called for.

All in all, data science is emerging into the mainstream software development projects. However, data science entails only a small portion of overall work and

the role of a data scientist is often clearly identifiable in the development team. Nevertheless, it is likely that the processes of data science will continue to draw practices from software development. There is a level of uncertainty that is an inherent trait of data science. Hence all processes suitable in the context of software development do not necessarily apply to data science work. Instead the processes themselves should evolve to be able to take data science components with built-in uncertainty into account as part of the process.

There is clearly need for further research on the nature of data science project methodology. First of all, the results found in this paper should be validated with larger source material. In addition, it would be interesting to further test different project methodology approaches in a data science environment. These results already indicate that data science is increasingly using development processes to guide the work and rely on experimentation and multidisciplinary team work.

Acknowledgments

The authors wish to thank the professionals who provided their time and experience for our interviews. This study would not have been possible without them.

References

1. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., Zimmermann, T.: Software engineering for machine learning: A case study. In: IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice (2019)
2. Angée, S., Lozano-Argel, S.I., Montoya-Munera, E.N., Ospina-Arango, J.D., Tabares-Betancur, M.S.: Towards an improved ASUM-DM process methodology for cross-disciplinary multi-organization big data & analytics projects. In: International Conference on Knowledge Management in Organizations (2018)
3. Azevedo, A., Santos, M.F.: KDD, SEMMA and CRISP-DM: A parallel overview. In: IADIS European Conference on Data Mining (2008)
4. Brachman, R.J., Anand, T.: The process of knowledge discovery in databases: A first sketch. In: AAAI Workshop on Knowledge Discovery in Databases (1994)
5. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology **3**(2) (2006)
6. Budde, R., Kautz, K., Kuhlenkamp, K., Züllighoven, H.: What is prototyping? In: Prototyping. Springer (1992)
7. Grady, N.W.: KDD meets big data. In: IEEE International Conference on Big Data (2016)
8. Grady, N.W., Payne, J.A., Parker, H.: Agile big data analytics: AnalyticsOps for data science. In: IEEE International Conference on Big Data (2017)
9. Hill, C., Bellamy, R., Erickson, T., Burnett, M.: Trials and tribulations of developers of intelligent systems: A field study. In: IEEE Symposium on Visual Languages and Human-Centric Computing (2016)
10. Kandel, S., Paepcke, A., Hellerstein, J.M., Heer, J.: Enterprise data analysis and visualization: An interview study. IEEE Transactions on Visualization and Computer Graphics **18**(12) (2012)

11. Kim, M., Zimmermann, T., DeLine, R., Begel, A.: The emerging role of data scientists on software development teams. In: IEEE/ACM International Conference on Software Engineering (2016)
12. Kim, M., Zimmermann, T., DeLine, R., Begel, A.: Data scientists in software teams: State of the art and challenges. Transactions on Software Engineering **44** (2018)
13. Piatetsky, G.: CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDnuggets (2014), retrieved June 2020 from `https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html`
14. Ries, E.: The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses. Currency (2011)
15. Runeson, P., Höst, M.: Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering **14**(2) (2008)
16. Safhi, H.M., Frikh, B., Hirchoua, B., Ouhbi, B., Khalil, I.: Data intelligence in the context of big data: A survey. Journal of Mobile Multimedia **13**(1&2) (2017)
17. Saltz, J., Shamshurin, I., Connors, C.: Predicting data science sociotechnical execution challenges by categorizing data science projects. Journal of the Association for Information Science and Technology **68** (2017)
18. Saltz, J.S.: The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In: IEEE International Conference on Big Data (2015)
19. Saltz, J., Hotz, N., Wild, D., Stirling, K.: Exploring project management methodologies used within data science teams. In: Americas Conference on Information Systems (2018)
20. Saltz, J.S., Shamshurin, I.: Exploring the process of doing data science via an ethnographic study of a media advertising company. In: IEEE International Conference on Big Data (2015)
21. Saltz, J.S., Shamshurin, I.: Big data team process methodologies: A literature review and the identification of key factors for a project's success. In: IEEE International Conference on Big Data (2016)
22. Schmidt, C., Sun, W.N.: Synthesizing agile and knowledge discovery: Case study results. Journal of Computer Information Systems **58**(2) (2018)
23. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D.: Hidden technical debt in machine learning systems. In: Advances in Neural Information Processing Systems (2015)
24. Shafique, U., Qaiser, H.: A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research **12** (2014)
25. Terho, H., Suonsyrjä, S., Systä, K., Mikkonen, T.: Understanding the relations between iterative cycles in software engineering. In: Hawaii International Conference on System Sciences (2017)
26. Wirth, R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. In: International Conference on the Practical Applications of Knowledge Discovery and Data Mining (2000)
27. Yin, R.K.: Case Study Research: Design and Methods. SAGE Publications, 5th edn. (2013)